nature ecology & evolution

ARTICLES

PUBLISHED: 13 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0020

Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life

Dahiana Arcila[1,2], Guillermo Ortí[1], Richard Vari[2,†], Jonathan W. Armbruster[3], Melanie L. J. Stiassny[4], Kyung D. Ko[1], Mark H. Sabaj[5], John Lundberg[5], Liam J. Revell[6] and Ricardo Betancur-R.[2,7*]

Much progress has been achieved in disentangling evolutionary relationships among species in the tree of life, but some taxonomic groups remain difficult to resolve despite increasing availability of genome-scale data sets. Here we present a practical approach to studying ancient divergences in the face of high levels of conflict, based on explicit gene genealogy interrogation (GGI). We show its efficacy in resolving the controversial relationships within the largest freshwater fish radiation (Otophysi) based on newly generated DNA sequences for 1,051 loci from 225 species. Initial results using a suite of standard methodologies revealed conflicting phylogenetic signal, which supports ten alternative evolutionary histories among early otophysan lineages. By contrast, GGI revealed that the vast majority of gene genealogies supports a single tree topology grounded on morphology that was not obtained by previous molecular studies. We also reanalysed published data sets for exemplary groups with recalcitrant resolution to assess the power of this approach. GGI supports the notion that ctenophores are the earliest-branching animal lineage, and adds insight into relationships within clades of yeasts, birds and mammals. GGI opens up a promising avenue to account for incompatible signals in large data sets and to discern between estimation error and actual biological conflict explaining gene tree discordance.

The advent of genomic approaches is delivering unprecedented amounts of sequence data from non-model organisms, sparking enthusiasm and heightening expectations about the resolution of ancient divergences in the tree of life[1]. Substantial controversy persists, however, concerning the best way to analyse genome-wide data sets, especially for taxonomic groups shown to be recalcitrant to phylogenetic resolution[2–6]. The conventional concatenation approach combines all gene alignments into a single data set or supermatrix prior to phylogenetic analysis. However, theory and simulations indicate that concatenation methods can yield misleading results when gene tree conflict is high, owing to incomplete lineage sorting (ILS)[7–11].

In the past decade, the field of molecular phylogenetics has shifted from concatenation methods to employing an increasingly diverse collection of multi-species coalescent approaches to account for ILS[10]. It is theoretically sound to use methods that model coalescent variance, particularly those that integrate over gene tree uncertainty in a Bayesian framework[10,12]. Yet, full parametric co-estimation of gene trees and species trees is not currently scalable to large, genome-wide data sets, which are instead analysed by reconciling a collection of pre-estimated individual gene trees under the coalescent. A major assumption of these 'summary'[13] or 'shortcut'[14] coalescent methods is that individual gene trees accurately depict the genealogical history of fragments of the genome that independently segregate (coalescent genes, or c-genes). To meet this theoretical challenge with empirical data sets, practitioners of phylogenetics have been trapped between two undesirable extremes. On one end, the analysis of short, recombination-free genes (consisting of a few hundred sites) are error-prone due to limited signal-to-noise content[2,3,14–17]. On the other extreme, long genes or full-length transcripts with thousands of sites harbor more phylogenetic information, reducing (but not necessarily removing) stochastic error[2,18]. Longer genes, however, are more likely to carry past recombination events, violating the assumption of a single genealogical history[8]. Both situations lead to statistical inconsistency under the multi-species coalescent, and these limitations have recently spurred heated debates over the merits of coalescent approaches for the analysis of ancient divergences[2,4,16,19–21]. As a consequence, various procedures have been proposed to mitigate gene tree estimation error, including binning short gene alignments to augment information content[16,17], selecting subsets of highly informative genes[5], or simply bypassing gene tree estimation by evaluating unlinked single-site data to infer quartet trees, subsequently combined into a species tree via quartet amalgamation[13,22]. Currently, no consensus to solve this conundrum has emerged, some of the proposed solutions have raised controversy[13,19,23], and ambiguity persists when different methods do not converge on a unique result[6,24].

Here, we present a phylogenomic approach that efficiently extracts the genealogical signal from short c-genes by reducing the complexity of tree space on the basis of topological constraints. This method is similar to others that place priors on gene tree topologies[25], but is unique in that priors are set to test specific hypotheses directly. We show how this procedure resolves longstanding controversies using newly generated data for otophysan fishes and published data sets for other exemplary groups (metazoans, neoavian birds, eutherian mammals and yeasts). Otophysan fishes constitute the dominant group in freshwater habitats around the world, having experienced one of the nine major radiations among jawed

[1]Department of Biological Sciences, The George Washington University, 2023 G Street NW, Washington DC 20052, USA. [2]Department of Vertebrate Zoology, National Museum of Natural History Smithsonian Institution, PO Box 37012, MRC 159, Washington DC 20013, USA. [3]Department of Biological Sciences, Auburn University, Auburn, Alabama 36849, USA. [4]Department of Ichthyology, Division of Vertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA. [5]Department of Ichthyology, The Academy of Natural Sciences, 1900 Benjamin Franklin Parkway, Philadelphia, Pennsylvania 19103, USA. [6]Department of Biology, University of Massachusetts Boston, Boston, Massachusetts 02125, USA. [7]Department of Biology, University of Puerto Rico – Río Piedras, PO Box 23360, San Juan, Puerto Rico. [†]Deceased. *e-mail: betanri@fishphylogeny.org

vertebrates[26]. The clade, comparable in diversity to birds, consists of more than 10,000 species arrayed into 77 families, 7 suborders, and 4 orders (Cypriniformes, Characiformes, Siluriformes and Gymnotiformes). Otophysans include the well-studied model species (zebrafish, *Danio rerio*), carps, minnows, characins (for example, tetras and piranhas), knifefishes (such as the electric eel) and catfishes. For the past three decades, the most widely accepted hypothesis of relationships among otophysan orders has been based on an exemplary morphological analysis (hereafter referred to as H_0)[27]. Molecular studies, in contrast, have produced conflicting phylogenetic results that differ from the null morphological tree[28–31], disagreeing about the interrelationships of all major groups, with some even failing to resolve the order Characiformes as monophyletic despite strong morphological support (Fig. 1 and Supplementary Table 1).

To address this challenging phylogenetic question, we collected genome-wide sequence data from 1,051 exons using target capture and Illumina sequencing for 225 species representing all major otophysan lineages (Supplementary Table 2). Exons targeted for this study were chosen from genome comparisons to select single-copy short sequences (with an average length of 200 bp), while avoiding long stretches of DNA to minimize recombination. Analyses of complete data sets, smaller subsets and individual gene fragments using a range of standard approaches designed to minimize conditions that may lead to systematic error failed to provide compelling support for a single phylogenetic hypothesis, suggesting that choice of method (concatenation or species trees), data subset (for example strong signal, conserved genes, and so on) or data type (DNA or protein sequences) strongly influences the outcome (Fig. 1). In this case, far from settling the dispute, best practice methodologies aimed at minimizing systematic error in phylogenomics seemed to exacerbate it — neither concatenation nor species tree methods, nor DNA-based or protein-based analyses, converge on a single topology. To gain additional insight, we developed an analytical approach based on topology tests that gauges the strength of phylogenetic signal contained in each gene alignment in favour of alternative hypotheses. By constraining gene-tree space to a small number of relevant options (15 in this case; Fig.1), this approach overcomes gene tree estimation error to reveal overwhelming evidence favouring H_0. To further assess the utility and performance of this approach, we re-examined published data sets for other groups with controversial phylogenetic relationships.

Genealogical signal of exon markers at different scales

Before inspecting incongruence among concatenation and species tree methods in regard to the central hypothesis being investigated (the interrelationships of otophysan lineages), we assessed the collective performance of exon markers in multi-locus analyses and the extent of estimation error for individual gene trees by: (i) evaluating support for uncontroversial groups (otophysan orders, suborders and families) that are independent of the central hypothesis (Supplementary Fig. 1); (ii) comparing tree space dispersion plots using multidimensional scaling (MDS) based on unweighted Robinson–Foulds distances[32] (Supplementary Fig. 2A); and (iii) estimating average support values across all clades in the corresponding trees (Supplementary Fig. 2B). The first test is a proxy for phylogenetic accuracy (the probability of resolving undisputed groups), whereas the latter two measure phylogenetic precision (the deviation of estimates in tree space and robustness of inferences).

Individual gene trees were estimated using standard partitioned maximum likelihood (ML) and Bayesian methods, whereas multi-locus analyses explored a large number of alternative approaches either involving concatenation or species tree methods, applied to multiple data sets (complete data or subsets filtered by properties) and data types (DNA and protein sequences; Supplementary Table 2) to account for potential systematic error due to base compositional biases[33,34]. For multi-locus methods, resolution of expected taxonomic groups of otophysans is almost unanimously obtained with high confidence (Supplementary Fig. 1). The resulting multi-locus trees are well supported (with an average support of 79.1%; Supplementary Fig. 2B) and appear tightly clustered in tree space (Supplementary Fig. 2A), suggesting high phylogenetic precision. These results indicate that, collectively, our exon markers contain strong phylogenetic signal at different evolutionary scales, and seem resilient to specific assumptions underlying each method.

By contrast, individual gene trees perform poorly both in terms of accuracy and precision, almost always failing to resolve undisputed groups (Supplementary Fig. 1), displaying topological distances in tree space that are orders of magnitude greater than those of multi-locus phylogenies (Supplementary Fig. S2A), and resulting in poorly supported clades (with an average support of 24.8%; Supplementary Fig. 2B). This result is not unexpected given that the average length of exons in our data set is 200 bp or 67 amino acids. Although short c-genes have the benefit of minimizing the risk of recombination, these results indicate that gene tree error is extensive.

Incongruence between concatenation and species trees

Despite the ability of multi-locus methods to resolve undisputed clades, the branching order of major otophysan lineages receives equivocal support (Fig. 1). We designed and implemented 45 different analyses of multi-locus data based on commonly applied concatenation and coalescent methods (Supplementary Table 2), implementing several criteria to minimize systematic error, and obtained support for 10 out of 15 possible topologies (Fig. 1). The distribution of results is decidedly uneven, with most concatenation methods supporting topology H_{a01} and most species tree methods supporting topology H_{a02}. Variants of both approaches also support other topologies, and H_0 ranks second or third in frequency (seven analyses support both H_0 and H_{a02}). No individual gene tree resolves any of these alternatives, confirming a high degree of estimation error based on single loci. These results suggest that in-depth exploration of phylogenomic data sets using alternative methods reflecting widely accepted best-practice criteria will reveal high levels of incongruence that is not easily integrated with current methodology to unambiguously support a single phylogeny[6,24]. Even more worrisome is the observation that conflicting topologies often receive strong bootstrap support, especially those resulting from concatenation analyses (Figs 1–3). We suggest that averaged support values from trees inferred from alternative analyses and data subsets (Supplementary Fig. 3) may provide a more realistic way to reflect nodal support and confidence in phylogenomics, while also accounting for incongruence inherent to data set type or method.

A method to overcome gene tree error

Instead of using error-prone gene trees as input for coalescent analyses, we devised 'gene genealogy interrogation' (GGI), an approach based on topology tests to identify the genealogical history, among a set of predefined alternatives, that each gene supports with highest probability. To establish the ranking of alternative trees and their probabilities, GGI implements constrained ML searches to optimize site likelihood scores for each gene alignment under each hypothesis. The method is based on the approximately unbiased (AU) topology test[35], which uses multi-scale bootstrapping techniques and can be applied to simultaneous comparisons of multiple trees. GGI is designed to address one phylogenetic problem at a time by defining a set of alternative hypotheses. If gene tree error is suspected to be a major source of conflict in other parts of the tree, then new GGI tests must be conducted.

We applied GGI to test the central hypothesis of otophysan relationships and examined all possible unrooted topologies for five lineages (Fig. 1), conducting a total of 31,530 constrained ML searches

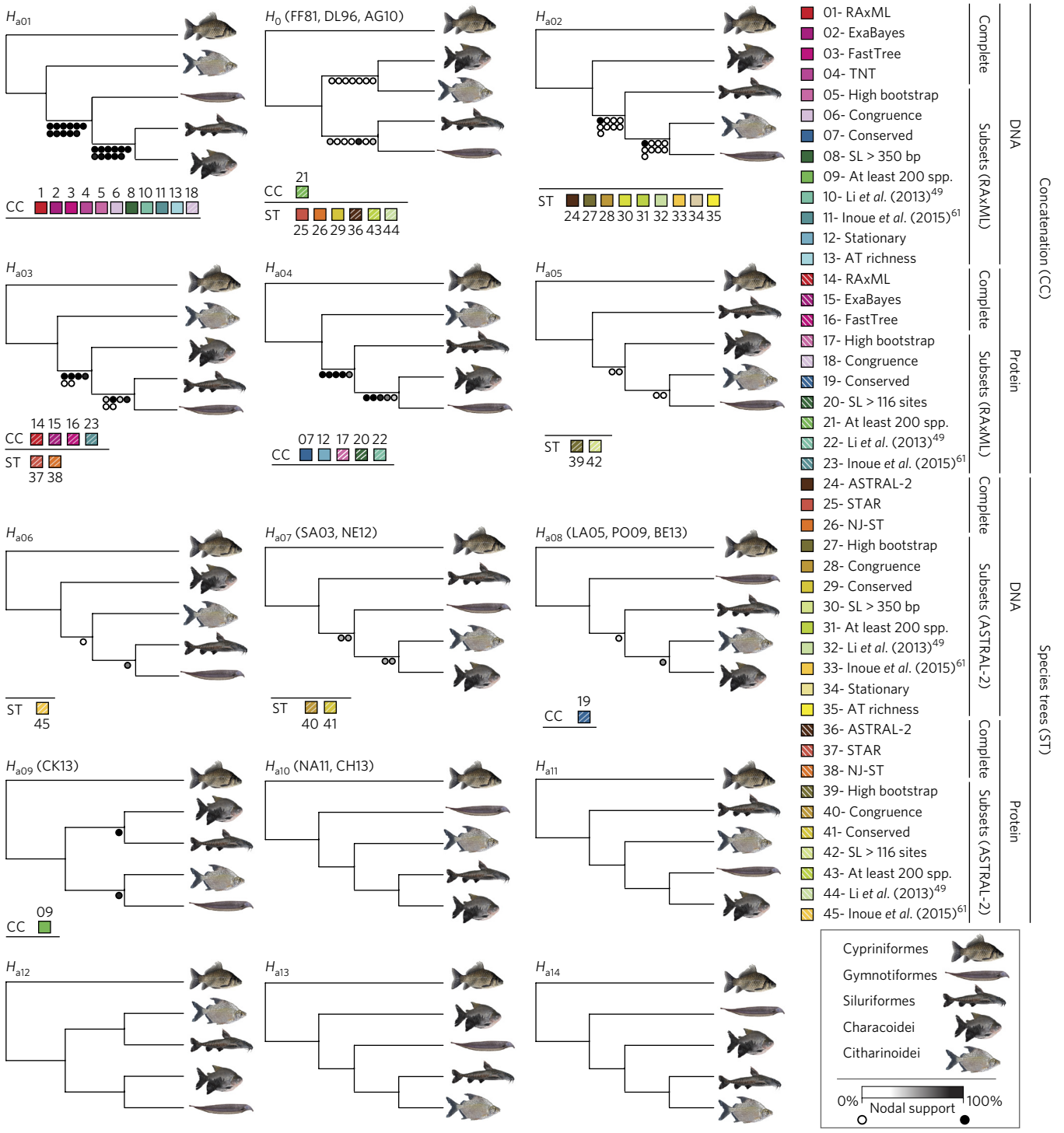

Figure 1 | Null morphological hypothesis (H_0) and all 14 possible alternative trees for the five major lineages in Otophysi. Previous studies supporting each hypothesis are listed in parenthesis above the corresponding tree: FF81[27], DL96, SA03, NE12, LA05, PO09, BE13, NA11, CH13 and CK13 (see Supplementary Table 1 for citation abbreviations). A total of 45 analytical approaches (see Methods) were applied to our data set of 1,051 loci, which collectively resolve 10 different trees. Analyses supporting each tree are listed under the corresponding trees (squares) and circles below branches indicate clade support for each method (ExaBayes: posterior probabilities; all other analyses: bootstrap values), following the same order (left to right, top to bottom) of coloured squares. All concatenation and species tree methods applied to data subsets use RAxML and ASTRAL-2, respectively. The order Characiformes (not indicated in figure), comprising the suborders Citharinoidei plus Characoidei, is monophyletic in 3 of 15 topologies. Fish image credits M.H.S.

(15 topologies for each of 1,051 gene trees, based on protein or DNA sequences). Here, each alternative hypothesis is defined by a different set of phylogenetic 'backbone' relationships between major lineages. In each optimization, we constrained each of the five major subclades to be monophyletic (see below and the Supplementary Information), but we imposed no other constraint with regard to relationships within each subclade, nor with respect to branch lengths nor model parameters. More than twice as many topology tests found that hypothesis H_0 was supported with the highest probability, for both DNA (495 loci) and protein (314 loci) data sets, compared to the

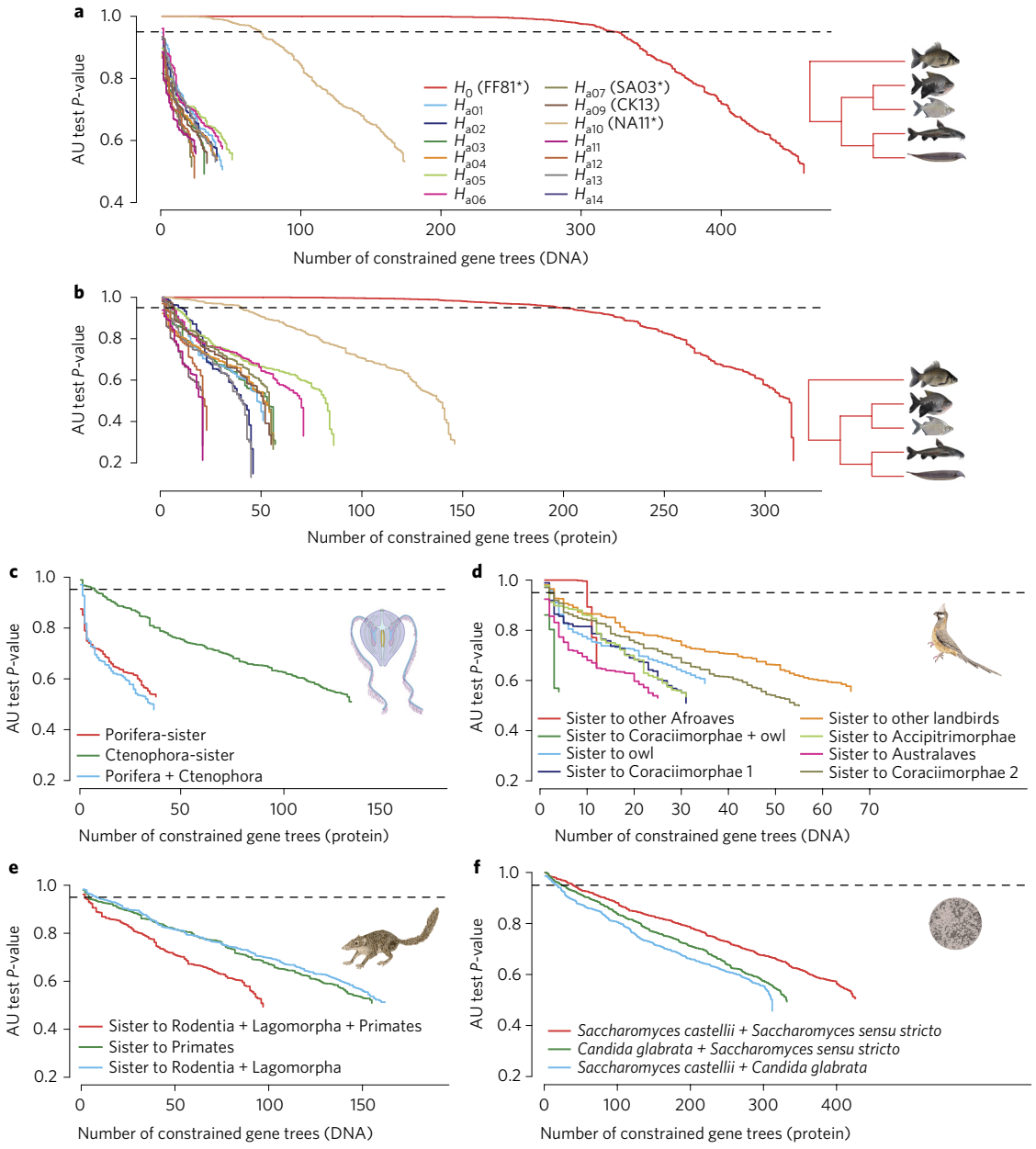

Figure 2 | Gene genealogy interrogation (GGI) applied to phylogenomic data sets to test alternative hypotheses. Lines represent the cumulative number of genes (*x* axes) supporting each hypothesis with highest probability (rank 1) and their associated *P*-values (*y* axes) according to the approximately unbiased (AU) topology test. Values above the dashed line indicate all rank 1 hypotheses that are significantly better than the alternatives ($P < 0.05$), whereas those below the dashed line are also rank 1 but without statistical significance. **a**,**b**, Otophysi (*see Supplementary Table 1). **c**, Metazoans. **d**, Neoaves (mousebird). **e**, Eutherian mammals (tree shrew). **f**, Yeast phylogeny.

second-best hypothesis (H_{a10}) with 174 and 146 tests in favour, respectively. This difference increased to 5-fold (325 versus 69 for DNA and 197 versus 39 for protein) for tests results where the best hypothesis (H_0) is significantly better ($P < 0.05$) than the second ranked hypothesis (H_{a10}). All alternative topologies received negligible support (Fig. 2a,b and Supplementary Table 3). Interestingly, both DNA- (Fig. 2a) and protein-based (Fig. 2b) GGI analyses produced similar results, suggesting that non-stationarity at the DNA level is not a significant systematic bias compromising the topology tests.

We acknowledge that while monophyly of the five major otophysan groups (subclades) is supported with high confidence by multiple lines of evidence at the species-tree level (morphology, mitochondrial DNA, multi-locus nuclear DNA and genomics[23,28-31]; Supplementary Fig. 1), it is possible that deep coalescences could result in particular gene histories that display non-monophyly of one or more subclades. This possibility, however, is highly unlikely because internal branches subtending subclades span 20–65 million years of evolution (Supplementary Table 7), providing ample time for the vast majority of gene genealogies to achieve reciprocal monophyly (see Supplementary Information). Thus, rare instances of deep coalescences for some genes resulting in the non-monophyly of subclades are unlikely to introduce biases beyond stochastic error into the GGI approach. We discuss this in detail in the Supplementary Information, and provide additional tests with a modified GGI procedure that relaxes the assumption of subclade monophyly by using a combination of unconstrained and constrained gene trees as input for summary coalescent analyses.

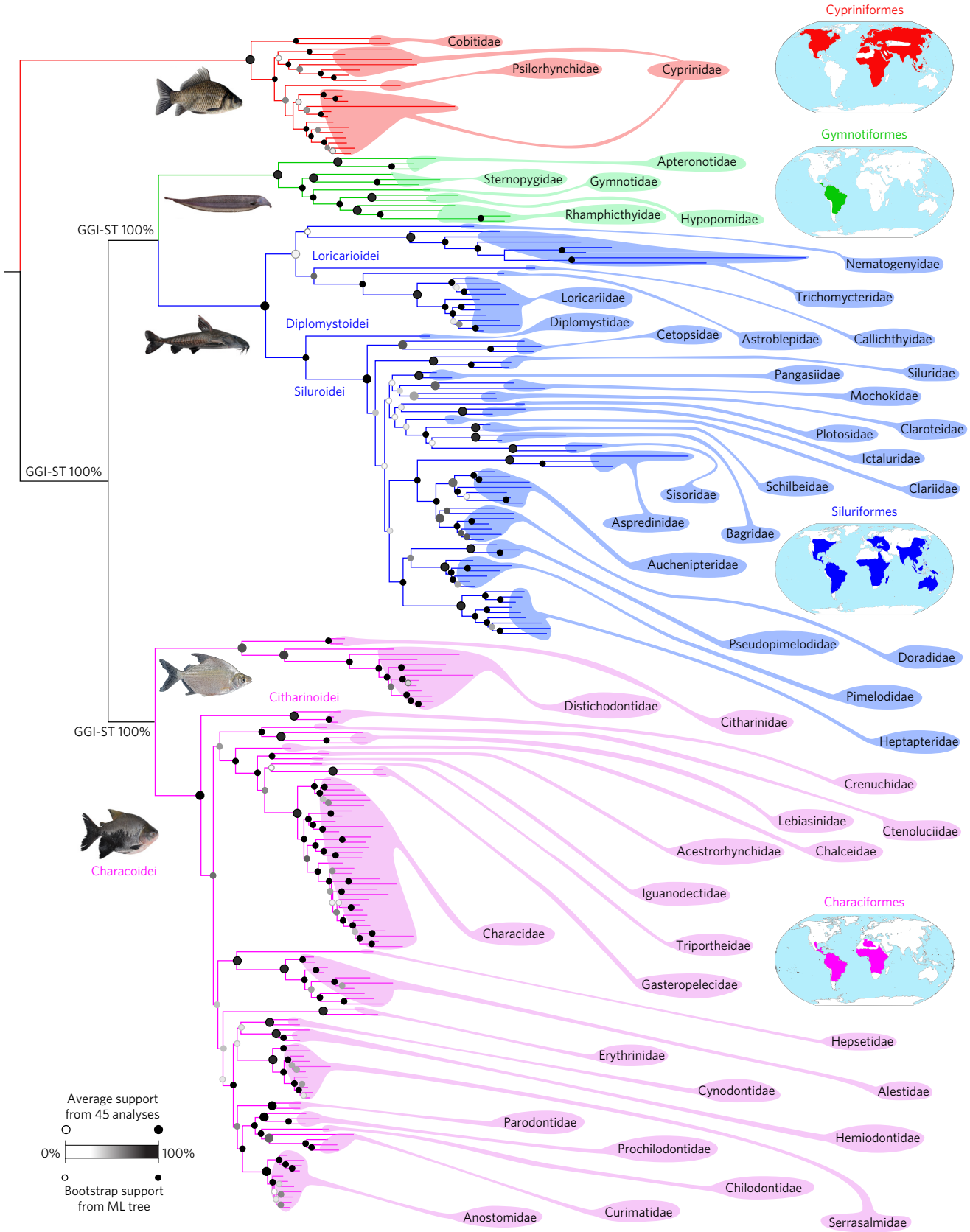

Figure 3 | Otophysan phylogeny based on the concatenation analysis of 231 protein genes with minimal missing information. This preferred topology (and branch lengths) is based on RAxML analysis 21, which is largely congruent with the GGI-based species tree topology obtained with STAR and ASTRAL-2 (Supplementary Fig. 3). Support values are based on the average obtained from 45 alternative analyses (large circles), the bootstrap support obtained from the ML tree (small circles), or the bootstrap values of the GGI-based species tree analyses using ASTRAL-2 (denoted as GGI-ST; key nodes tested only). Marine otophysan lineages are not depicted in the maps.

ILS is not the main problem

Coalescent theory predicts that phylogenetic histories of lineages evolving under a combination of short internal branches and large effective population sizes are prone to high incidence of ILS[8]. It has been demonstrated that for five or more lineages such conditions can generate gene trees with topologies that differ from the underlying species phylogeny with highest probability[36,37]. When the evolutionary history of a clade falls within this so-called anomaly zone[8], simply adopting the most frequent gene tree as a surrogate for the species phylogeny (the democratic vote procedure) is positively misleading.

To account for this possibility, as the genuine backbone tree of inter-ordinal relationships must be 1 of 15 possibilities (enumeration of all possibilities for an unrooted tree of five taxa), we used the GGI trees selected by the topology tests (the preferred constrained gene trees optimized by ML) as input for summary coalescent analyses. For this test we employed both DNA- and protein-based trees in combination with two different species-tree methods. We also applied two alternative approaches for sampling GGI trees, one using all rank 1 trees (complete data with 1,051 genes) and another using only the set of rank 1 trees that are significantly better than the alternatives ($P < 0.05$; a subset of 397 DNA trees and 275 protein trees; Supplementary Table 3). Of the eight species-tree analyses conducted, all converged on the H_0 tree, with each backbone node receiving 100% bootstrap support. Finally, an adapted version of the GGI-based coalescent method that uses constrained topologies in combination with unconstrained gene trees also supports the H_0 tree (Supplementary Information).

Our results suggest that the evolutionary history of major otophysan lineages is not trapped in the anomaly zone. In fact, these analyses identify only a minor proportion of gene trees that are significantly discordant with the inferred species phylogeny (17.7–28.4%, most supporting H_{a10}), suggesting that other sources of error rather than ILS are likely the main cause of incongruence. Gene tree estimation error may be biasing summary coalescent approaches, but the causes for discrepancy between coalescent and concatenation results are unclear. For two hypotheses (H_0 and H_{a03}), some concatenation and species tree methods converge, but more often they seem to produce non-overlapping sets of results (Fig. 1). We were unable to isolate any single factor as the principal explanation for discordance in multi-locus analysis. Possibilities include the combination of slight model misspecifications interacting in analyses of large data sets and amplifying systematic biases, or processes such as horizontal gene transfer or duplication/extinction affecting some of the sampled genes[38]. What is perhaps most surprising is the observation that the most common topology from concatenation is incongruent with our GGI tree, even in the absence of evidence for substantial ILS. An investigation of factors that could account for this pattern would be a fruitful subject of future theoretical and analytical studies. In summary, the coalescent analyses using GGI trees resolve with high confidence the branching order of major otophysan groups (Supplementary Fig. 3), a result that is fully congruent with the morphological hypothesis (H_0)[27], thereby reconciling a long history of molecular and morphological conflict.

Addressing other recalcitrant clades with GGI

To test the generality of the GGI approach, we conducted additional tests using published phylogenomic data sets for distantly related groups with controversial resolution in the tree of life (Supplementary Table 4). We chose four emblematic phylogenetic questions that have recently received substantial attention: (i) the position of sponges and ctenophores (comb jellies) at the base of the animal (metazoan) tree[39–42]; (ii) the relationship of the mousebird to other lineages within the Neoaves radiation[43–45]; (iii) the position of the tree shrew (Scandentia) among eutherian mammals[20,46]; and (iv) the relative placement of *Candida glabrata*, *Saccharomyces castellii* and *Saccharomyces sensu stricto* in the yeast phylogeny[4].

Two contrasting patterns emerge from these tests (Fig. 2c–f and Supplementary Table 3). First, as in the case of otophysans, the metazoan data set provides strong differential support in favour of a single topology. While the traditional view has been that sponges are the first branching lineage in the animal tree, most recent phylogenomic studies support the so-called Ctenophora-sister hypothesis that places comb jellies as the sister group to all other animals (refs. [34–36] but see ref. [33]). In agreement with recent genomic evidence, the vast majority of GGI trees favour the latter hypothesis (Fig. 2c). Second, in none of the other three groups (yeasts, mammals and Neoaves) does GGI select a particular tree topology over another with overwhelming support (Fig. 2d–f), indicating that genealogical conflict in these groups is substantial.

For metazoans, yeast and mammals, we test hypotheses involving only four lineages, implying that only three possible topologies need to be considered. Because rooted three-taxon (or unrooted four-taxon) species trees are free from anomalies under the coalescent[8,36,37], the most frequent gene tree topology in these cases may be interpreted as the species phylogeny (assuming subclade monophyly in individual gene trees is undisrupted by deep coalescences; see Supplementary Information). A topology supporting the clade *Saccharomyces castellii* + *Saccharomyces sensu stricto* is more frequently favoured (426 genes) than the two other alternatives (332 and 312 respectively) based on the yeast data set (Fig. 2f). This result is consistent with the gene tree frequencies originally reported[4]. For the mammalian data set (Fig. 2e), the GGI results prefer one of two competing hypotheses, albeit by a small difference: 155 genes place the tree shrew (Scandentia) as sister to primates, as claimed by the original study, whereas 165 genes place it as sister to a clade including Rodentia plus Lagomorpha, in agreement with another reanalysis of this data set[3,20]. The placement of the mousebird among major neoavian lineages is a six-taxon problem that entails tests for 105 possible topologies (an analysis beyond the scope of this study). Our preliminary GGI analyses did, however, provide a test among eight competing hypotheses[43,44], favouring with statistical significance the position of the mousebird as sister to other Afroaves[44]. For this case, high levels of gene tree discordance have been attributed to pervasive ILS during the early diversification of Neoaves[44], requiring the set of 105 topology tests for GGI-based coalescent analyses.

Perspective

Our GGI method provides a promising avenue to address difficult phylogenetic problems by accounting for gene tree estimation error through topology tests. The method interrogates individual gene partitions by constraining tree space to evaluate the relative support for specific hypotheses. This principle has been applied by other methods but without *a priori* references[25], or using Bayesian applications that do not scale up to genome-level data sets[10,12]. Thus, GGI has the favourable property of avoiding potential pitfalls inherent to concatenation and many other species tree approaches.

For our otophysan data set, GGI resolves a longstanding question in fish systematics and provides unambiguous support for the null morphological tree[27]. This reconciliation has remained elusive in most previous molecular studies[28–30], including another recent investigation using genome-wide data[31], and our in-depth analysis using standard concatenation and coalescent approaches. Our result is consistent with the monophyly of Characiformes and with a single evolutionary origin of electric organs within Otophysi (Supplementary Fig. 3), conditions shared by only catfishes and knifefishes, and strongly supported on morphological grounds.

Confidence in the selection of a preferred hypothesis provided by the AU test mitigates sampling error in tree estimation arising from limited signal in small gene partitions (that is, data sets composed of short gene fragments that are otherwise free of recombination),

and avoids systematic biases with additive effects in large data sets. For cases where the main hypothesis can be defined in terms of an unrooted four-taxon statement (such as metazoans, mammals and yeasts), our GGI approach is expected to meet the statistical consistency of gene-tree 'democratic vote', even if severely affected by ILS. For problems involving five or more lineages (for example, otophysans and birds), we propose and apply a pipeline whereby we first estimate a set of plausible gene trees under our alternative hypotheses, rank them for each gene, and then use the highest ranked gene trees (under different criteria) as input for summary species-tree analysis (GGI-based species tree). For cases in which deep coalescences may result in the violation of the assumption of subclade monophyly imposed by the topological constraints (thereby making the assignment to specific *n*-taxon statements difficult), we apply a modified version of the GGI-based coalescent procedure that uses a mixture of constrained and unconstrained gene trees (Supplementary Information). Tree distributions obtained with GGI, combined with the coalescent analyses, may prove useful for a broad class of data sets as a practical option to resolve stubbornly ambiguous clades in the tree of life.

In conclusion, the effect of sampling error in gene tree estimation is often overlooked when implementing summary coalescent approaches to resolve ancient divergences and/or recalcitrant clades in the Tree of Life using genome-wide data[3,15,16]. Our study shows that gene genealogy interrogation is a useful tool to distinguish between estimation error and actual biological conflict in explaining gene tree discordance, ultimately improving phylogenetic reconstructions of complex events such as the early diversification of otophysan fishes. We acknowledge that correct interpretation of the signal of gene tree discordance requires holistic models accounting for all biological processes that affect phylogenetic reconstruction (such as ILS, paralogy and reticulation)[38,47]. Until such models become available and efficient enough to synthesize large numbers of gene trees, GGI is a promising way forward because it provides explicit tests for gene tree incongruence around hard-to-resolve nodes, increasing our ability to infer organismal phylogeny.

Methods

A flowchart of the experimental design and methodological approaches used is shown in Supplementary Fig. 4. Details of the pilot study are explained in the Supplementary Information. Databases are archived in Zenodo (http://dx.doi.org/10.5281/zenodo.51603). We first conducted a pilot experiment to sequence 3,957 orthologous exons using target enrichment (TE)[48] and Illumina (Supplementary Table 5). We selected exons by screening the zebrafish and medaka genomes for single-copy, slowly evolving genes[49]. Probes designed using zebrafish sequences were hybridized with the genomic DNA of 14 species encompassing the diversity of ray-finned fishes. We then chose a subset of single-copy exons exhibiting highest capture efficiency among otophysans, and designed a new probe set based on sequences from four otophysans and five outgroups obtained in our experiment. We used these markers to collect 1,051 protein-coding sequences for 225 species representing 53 (of 77) families (279,012 DNA or 92,901 protein sites). We estimated DNA- and protein-based gene trees using partitioned maximum likelihood (ML) and Bayesian approaches.
To investigate incongruence and to identify the set of possible evolutionary histories of major otophysan lineages, we conducted a total of 45 different multi-locus analyses. These comprised concatenation (23 analyses) or coalescent-based methods (22 analyses); using either DNA (25 analyses) or protein (20 analyses) data sets; including complete data (13 analyses) or subsets of ~200 genes filtered following recommended criteria (32 analyses). Properties for subset selection include slowly evolving genes, strong phylogenetic signal, AT-richness, stationarity, and data completeness. Given their uncontroversial placement as first branching clade in Otophysi[28–31,50–54], we used cypriniform taxa as outgroups in all analyses. We conducted GGI for the set of 15 possible trees defining relationships among major otophysan lineages by constraining their monophyly and by assessing the ranking of alternative topologies for each gene via the AU topology test using CONSEL v0.1[55]. Finally, we conducted additional GGI analyses in other emblematic groups based on recently published phylogenomic data sets: metazoans (209 protein alignments[40]), Neoaves (259 DNA alignments[43]), eutherian mammals (414 DNA alignments[20,46]), and yeast (1,070 protein alignments[4]).

Genomic data collection (Otophysi). *Probe design.* A total of 1,041 target loci were selected from the pilot study (Supplementary Methods), and 21 markers extensively used by previous molecular studies were added to the marker set[56,57], including the mitochondrial COI gene for quality control (Supplementary Database 4). A new probe library was designed to capture the set of 1,041 slowly evolving exons based on sequences from nine species examined in the pilot study (*Pellona*, *Chanos*, *Kneria*, *Tanichthys*, *Danio*, *Apteronotus*, *Brustarius*, *Astyanax* and *Oryzias*). Probes for the remaining 21 markers were designed on the basis of sequences obtained from GenBank for 55 species representing major otophysan lineages. A total of 20,000 RNA baits (2× tiling) were synthesized by MYcroarray for the 1,062 marker set (Supplementary Database 5).

Taxonomic sampling. Tissue samples were collected from species that included 110 representative characiforms (12 from suborder Citharinoidei in 2 families, and 98 from suborder Characoidei in 21 families), 79 siluriforms (23 families) and 13 gymnotiforms (5 families). Because monophyly of Cypriniformes and its placement as the earliest branching otophysan lineage is uncontroversial, we only included 23 cypriniform species (4 families), and all were used as outgroups (Supplementary Table 1). In total, 10 samples yielded poor DNA quality; 6 others had to be excluded due to cross-contamination (detected by comparing COI sequences). The final taxonomic sampling consisted of 225 species representing 53 of the 77 valid families of Otophysi. Most samples sequenced include voucher specimens deposited in various museum collections (Supplementary Table 5).

Data collection and processing. For each sample, genomic DNA was extracted from fin or muscle tissue using a phenol-chloroform protocol in the Autogen platform. Library preparation, TE and Illumina sequencing (single-end) was outsourced to Rapid Genomics and followed the same protocols used for the pilot experiment. FASTQ files were trimmed using Geneious Pro v8.1 (http://www.geneious.com) with an error probability cutoff of 0.01. Contigs were assembled by mapping sequences against the zebrafish reference using the 'medium sensitivity' algorithm in Geneious with five iterations. The resulting contigs that assembled with <10× coverage and that were shorter than 75 bp were removed. Two loci with substantial amounts of missing data and nine loci producing more than one contig for at least one species after assembly also were excluded.

In summary, three consecutive steps were implemented to filter out putative cases of paralogy. (i) *In silico* screening of zebrafish and medaka genomes using reciprocal BLAST searches in EvolMarker[58] to select single-copy genes for the initial marker set (see pilot study in Supplementary Information). Although single-copy genes are defined on the basis of similarity thresholds, genes that share this property among distantly related genomes are probably orthologous. Gene duplications that take place in particular lineages may lead to the presence of in-paralogues that will not necessarily confound phylogenetic analysis of ancient divergences, but judicious exclusion of these may be warranted. (ii) Removal of 279 out of the initial 3957 loci used in the pilot study that produced two or more contig assemblies for at least one species. (iii) Removal of nine loci that resulted in two or more contigs for at least one species in the otophysan data set.

The final marker set consisted of 1,051 protein-coding genes (279,012 sites). The set of sequences obtained were aligned using MAFFT on a locus-by-locus basis. All alignments were visually inspected and edited to check for open reading frames. Seventy-one exon alignments had ambiguously aligned internal blocks that were removed to improve positional homology and to enable translation (Supplementary Databases 6 and 7). Alignments were translated to proteins using Translator X[59]. The final set of 1,051 exons were annotated using gene ontology (GO) in Blast2GO v3.2 (http://www.blast2go.org/), with a E-Value-Hit-Filter of 10^{-6}, an annotation cut-off of 55, and a GO weight of 5 (Supplementary Database 8).

Phylogenetic analysis and alternative data matrices. Forty-five different multi-locus analyses were conducted. These comprised concatenation (23 analyses) or coalescent-based species tree methods (22 analyses) that used either DNA (25 analyses) or protein sequence data (20 analyses), for the complete data set (13 analyses) or of data subsets of ~200 genes (32 analyses; Supplementary Table 2). Subsets of markers were selected based on criteria recommended by previous studies [2–4,34,49,60,61] as an attempt to minimize systematic biases and phylogenetic artefacts (Supplementary Database 9). The subsets (Fig. 1) were assembled following eight criteria (explained in detail in the Supplementary Information):

1. Gene trees with highest average bootstrap support (analyses 05, 17, 27 and 39). Following Salichos and Rokas[4], this subset includes 200 loci that resulted in gene trees harbouring the highest average bootstrap support (BS) values across all internodes (estimated with RAxML). Average BS values were estimated with the phylogenetic package Ape[62] using R[63]. The average BS values were 65% and 35% for the DNA and protein subsets, respectively. See details under 'Phylogenetic inference' (below).
2. Gene tree congruence (analyses 06, 18, 28 and 40). To reduce gene tree estimation error, a subset of 210 gene trees with the lowest average pairwise Robinson–Foulds (RF) distance was selected following recommendations and using scripts provided by Simmons *et al.*[3]. Selected gene trees based on DNA

and protein data sets had average RF distances of 0.70–0.81 and 0.96–0.90, respectively. Outlier gene trees were discarded by taking into account the number of shared terminals in pairwise comparisons.

3. Slowly evolving genes (analyses 07, 19, 29 and 41). The most conserved locus set was selected for phylogenetic analysis[2]. The 200 alignments with highest average identity (88–95% and 96–100% for DNA and protein alignments, respectively) were selected using Geneious.
4. Exons with longer sequences (analyses 08, 20, 29 and 42). This subset includes 205 locus alignments whose sequence length is greater than 350 nucleotides (60,225 sites) or 96 amino acids (28,907 sites). The underlying criterion is that longer exons harbour better signal-to-noise ratios that would minimize gene tree error.
5. Minimizing missing data (analyses 09, 21, 31 and 43). All single-locus alignments that had at least 200 species (out of 225) were included to minimize empty cells per taxon in the corresponding gene matrices, thus reducing the proportion of missing data[60]. A total of 231 loci were selected for both DNA (68,682 sites) and protein (22,441 sites) sequence sets.
6. Genes shared with other studies (analyses 11, 12, 22, 23, 32, 33, 44 and 45). Two subsets were assembled following recent studies that used exon-based phylogenomics in fishes and applied different criteria for marker selection (Li *et al.*[49]; Inoue *et al.*[61]). A total of 243 loci (60,147 sites) in common with Li *et al.*[49] and 175 loci (44,559 sites) in common with Inoue *et al.*[61] were selected.
7. Genes with minimal base compositional bias (analyses 12 and 34). This criterion seeks to minimize potentially misleading effects of base composition heterogeneity. We showed that mean disparity index (DI) estimated from all pairwise comparisons for each gene alignment provides a useful metric to rapidly assess the degree of compositional heterogeneity in multiple gene partitions[33]. A total of 200 loci (46,677 sites) with the lowest mean DI (0.0096–0.078) were selected using MEGA5[64].
8. Highest AT content (analyses 3 and 35). Romiguier *et al.*[34] showed that GC-rich genes result in higher levels of gene-tree error and incongruence relative to AT-rich loci. Percentage AT for each locus alignment was estimated using Geneious and a set of 200 loci (52,809 sites) with the highest AT content (49–60%) was selected.

Assessment of tree inference accuracy and precision. We conducted three different analyses to assess the collective performance of exon markers in multi-locus analyses and the extent of estimation error in individual gene trees. First, we gauged the power of the data to resolve and support taxonomic groups (otophysan orders, suborders, and families) that are undisputed in the literature and recognized on the basis of ample morphological and molecular evidence. These groups are independent from our central hypothesis tested. The presence of expected clades in multi-locus analyses and individual gene trees was assessed using the R package MonoPhy[65]. Twelve families represented by only one individual in this study were not tested (Supplementary Fig. 1). Second, we analysed discordance among 1,051 gene trees by graphically representing their dispersion in tree space in comparison with the 45 multi-locus trees. This test used multidimensional scaling (MDS) based on un-weighted Robinson–Foulds distances[32] as implemented in the TreeSetViz module in Mesquite[66]. The MDS analyses were conducted separately for DNA- and protein-based trees. Third, we estimated average support values across all nodes in the corresponding trees using the R package Ape.

Phylogenetic inference. *Concatenation-based analyses.* All alignments were concatenated into a single super-matrix for phylogenetic analysis based on the complete data set (1,051 loci) or on subsets described above (Supplementary Database 10). For all data sets, partitioned RAxML analyses (by gene and by codon position), were replicated 30 times and the best-scoring tree across searches was selected. DNA analyses used the GTRGAMMA model and protein analyses the PROTGAMMAWAG model in RAxML. Branch support was assessed using the rapid bootstrap algorithm with 300 replicates under the previous models; the collection of bootstrapped trees was used to draw bipartition frequencies onto the optimal tree. Additional unpartitioned analyses for the complete data sets (1,051 loci) were conducted using FastTree-2[67] under the GTR (DNA) and WAG (protein) models; FastTree local support values were estimated with the Shimodaira-Hasegawa test[35].

Bayesian analyses were run using ExaBayes v1.4.1[68] under the GTRGAMMA (DNA) and PROTGAMMAWAG (protein) models, with branch lengths linked across partitions. Two independent MCMC runs were conducted from random starting topologies sampling every 500 generations. ExaBayes runs continued until the termination condition of mean topological differences was less than 5% with at least 500,000 generations. Posterior distributions of trees were summarized using the 'consense' function with default burn-in. Convergence was assumed when all parameters had effective sampling sizes (ESS) greater than 200 estimated with Tracer v1.5[69]. In addition to model-based inference approaches, parsimony searches were performed for the complete nucleotide alignment in TNT v1.0[70]. The runs used the 'new technology' search option, with sectorial, ratchet and tree-fusing methodologies, with default parameters. To assess branch support, 100 bootstrap searches were performed via TBR branch swapping (summarized in a consensus tree).

Gene tree inference. Individual gene trees were inferred using RAxML and ExaBayes, as explained above. To assess performance in gene tree estimation between these two methods, we computed Robinson Foulds (RF) distances among each gene tree and a reference topology (estimated with the complete concatenated data set). RAxML produced gene-trees with smaller dispersion in tree topology relative to ExaBayes (smaller average RF distances); therefore, RAxML gene trees were used for downstream analyses.

Summary coalescent species-tree inference. Species-tree analyses were conducted for all data sets using ASTRAL-2 (Database S10). This method uses unrooted gene trees as input and maximizes the number of quartet trees shared between the gene trees and the species trees. ASTRAL-2 has been shown to outperform other summary methods under different levels of incomplete lineage sorting. To account for gene tree estimation uncertainty and to assess clade support, we used 100 RAxML bootstrapped gene trees for each gene (as described above) as input for ASTRAL-2. Additional summary coalescent analyses were performed using STAR and NJ-ST[71,72], as implemented in the STRAW server[73]; complete data sets only). All STAR and NJ-ST analyses were rooted using *Danio rerio*; all other cypriniform taxa were excluded. Details on assessment of tree inference accuracy and precision are given in the Supplementary Methods.

Gene genealogy interrogation (GGI). The GGI tests implemented require three major steps. First, we define a set of hypotheses to test: for our study, this includes the 15 possible topologies (Fig. 1) for the major lineages of otophysans (undisputed monophyletic groups: Cypriniformes, Gymnotiformes, Siluriformes, Characoidei, and Citharinoidei). Topological constraints enforcing these 15 hypotheses were defined to obtain 1,051 ML genes trees tree consistent with each hypothesis. Site likelihood scores for each tree were obtained with RAxML. Second, a topology test was conducted for each gene by statistically comparing the site likelihood scores of all 15 trees via the approximately unbiased (AU) test[35] as implemented in CONSEL v0.1[55]. The AU test uses multi-scale bootstrapping techniques and can be applied to simultaneous comparisons of multiples trees to estimate a *P*-value for each topology. Finally, trees were ranked according to the *P*-values and visualized using R plots supporting each hypothesis with highest probability. A tutorial for conducting all GGI steps using custom code is provided in the SI Text (Supplementary Databases 11 and 12).

Data sets analysed using GGI. In addition to the newly generated genomic set for otophysans (a five-taxon problem involving 15 possible topologies; Figs 1 and 2), four published data sets addressing controversial phylogenetic questions were analysed using GGI (Supplementary Table 3 and Database 13).

1. Metazoa (protein). A four-taxon problem involving three possible topologies (Fig. 2c). The metazoan data set analysed was compiled by Whelan *et al.*[40] and consists of 76 taxa and 209 genes. These authors assembled 25 alternative data sets, and this study examined their data set 12, which applies the most stringent filter for selection of orthologous loci (that it, 'certain' and 'uncertain' paralogues excluded[40]). It also comprises the broadest taxonomic sampling including distant outgroups such as fungi. Some studies assessing early metazoan relationships exclude distant outgroups to avoid potential artefacts caused by long-branch attraction[74]. However, this is not a concern in this study as GGI constrains the ingroup (animals) to be monophyletic.
2. Neoavian birds (DNA). A six-taxon problem involving 105 possible topologies, of which only eight competing hypotheses were assessed (Fig. 2d). This study examined the data set of Prum *et al.*[43], which includes 259 loci (consisting of exons and flanking introns) sequenced for 198 bird species. To reduce computational burden, this data set was subsampled to include a subclade in the Neoavian radiation where the mousebird is placed. The lineages sampled comprise Accipitrimorphae (7 species), Australaves (55 species), Coraciimorphae (23 species), owls (2 species), mousebirds (2 species) and one outgroup of the family Optisthocomidae.
3. Eutherian mammals (DNA). A four-taxon problem involving three possible topologies (Fig. 2e). The data set examined was originally assembled by Song *et al.*[75], consisting of 447 genes and 37 mammalian species. We used a recent correction of this data set[46], which relabeled two taxa inadvertently mislabeled in the original data set. We also excluded eight duplicate loci and 26 loci with misaligned sequences, following Springer *et al.*[20]. The data set examined consists of 413 genes.
4. Yeast (protein). A four-taxon problem involving three possible topologies (Fig. 2f). The yeast data set consists of 23 species and 1,070 exons assembled by Salichos and Rokas[4], with loci selected based on synteny and orthology information obtained from two genomic databases for yeasts.

Data availability. Data that support the findings of this study have been deposited in Zenodo (http://dx.doi.org/10.5281/zenodo.51603).

Received 10 June 2016; accepted 25 October 2016; published 13 January 2017

References

1. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798–804 (2003).
2. Betancur-R., R., Naylor, G. & Orti, G. Conserved genes, sampling error, and phylogenomic inference. *Syst. Biol.* **63,** 257–262 (2014).
3. Simmons, M. P., Sloan, D. B. & Gatesy, J. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Mol. Phylogenet. Evol.* **97,** 76–89 (2016).
4. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497,** 327–331 (2013).
5. Chen, M. Y., Liang, D. & Zhang, P. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. *Syst. Biol.* **64,** 1104–1120 (2015).
6. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22,** 225–231 (2006).
7. Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56,** 17–24 (2007).
8. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24,** 332–340 (2009).
9. Sen, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad. Sci. USA* **109,** 14942–14947 (2012).
10. Edwards, S. V., Liu, L. & Pearl, D. K. High-resolution species trees without concatenation. *Proc. Natl Acad. Sci. USA* **104,** 5936–5941 (2007).
11. Roch, S. & Steel, M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* **100C,** 56–62 (2014).
12. Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27,** 570–580 (2010).
13. Chou, J. *et al.* A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* **16,** S2 (2015).
14. Gatesy, J. & Springer, M. S. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.* **80,** 231–266 (2014).
15. Roch, S. & Warnow, T. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst. Biol.* **64,** 663–676 (2015).
16. Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346,** 1250463 (2014).
17. Bayzid, M. S., Mirarab, S., Boussau, B. & Warnow, T. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS ONE* **10,** e0129183 (2015).
18. Shen, X. X., Salichos, L. & Rokas, A. A genome-scale investigation of how sequence-, function-, and tree-based gene properties influence phylogenetic inference. *Genome Biol. Evol.* **8,** 2565–2580 (2016).
19. Liu, L. & Edwards, S. V. Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree". *Science* **350,** 171 (2015).
20. Springer, M. S. & Gatesy, J. The gene tree delusion. *Mol. Phylogenet. Evol.* **94,** 1–33 (2016).
21. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94,** 447–462 (2016).
22. Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30,** 3317–3324 (2014).
23. Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. Response to Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree". *Science* **350,** 171 (2015).
24. Posada, D. Phylogenomics for systematic biology. *Syst. Biol.* **65,** 353–356 (2016).
25. Wu, Y. C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: statistically informed gene tree error correction using species trees. *Syst. Biol.* **62,** 110–120 (2013).
26. Alfaro, M. E. *et al.* Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl Acad. Sci. USA* **106,** 13410–13414 (2009).
27. Fink, S. V. & Fink, W. L. Interrelationships of the ostariophysan fishes (Teleostei). *Zool. J. Linnean Soc.* **72,** 297–353 (1981).
28. Saitoh, K., Miya, M., Inoue, J. G., Ishiguro, N. B. & Nishida, M. Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. *J. Mol. Evol.* **56,** 464–472 (2003).
29. Nakatani, M., Miya, M., Mabuchi, K., Saitoh, K. & Nishida, M. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. *BMC Evol. Biol.* **11,** 177 (2011).
30. Chen, W. J., Lavoue, S. & Mayden, R. L. Evolutionary origin and early biogeography of otophysan fishes (Ostariophysi: Teleostei). *Evolution* **67,** 2218–2239 (2013).
31. Chakrabarty, P., McMahan, C., Fink, W., Stiassny, M. L. & Alfaro, M. in *ASIH – American Society of Ichthyologists and Herpetologists* (eds Crump, M. L. & Donnelly, M. A.) (2013).
32. Hillis, D. M., Heath, T. A. & St. John, K. Analysis and visualization of tree space. *Syst. Biol.* **54,** 471–482 (2005).
33. Betancur-R., R., Li, C., Munroe, T. A., Ballesteros, J. A. & Orti, G. Addressing gene-tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). *Syst. Biol.* **62,** 763–785 (2013).
34. Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N. & Douzery, E. J. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* **30,** 2134–2144 (2013).
35. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51,** 492–508 (2002).
36. Allman, E. S., Degnan, J. H. & Rhodes, J. A. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.* **62,** 833–862 (2010).
37. Degnan, J. H. Anomalous unrooted gene trees. *Syst. Biol.* **62,** 574–590 (2013).
38. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46,** 523–536 (1997).
39. Pisani, D. *et al.* Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl Acad. Sci. USA* **112,** 15402–15407 (2015).
40. Whelan, N. V., Kocot, K. M., Moroz, L. L. & Halanych, K. M. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl Acad. Sci. USA* **112,** 5773–5778 (2015).
41. Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* **45,** 371–395 (2014).
42. Ryan, J. F. *et al.* The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science* **342,** 1242592 (2013).
43. Prum, R. O. *et al.* A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526,** 569–573 (2015).
44. Suh, A., Smeds, L. & Ellegren, H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol.* **13,** e1002224 (2015).
45. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346,** 1320–1331 (2014).
46. Song, S., Liu, L., Edwards, S. V. & Wub, S. Correction for Song *et al.*, Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad. Sci. USA* **112,** E6079 (2015).
47. Hahn, M. W. & Nakhleh, L. Irrational exuberance for resolved species trees. *Evolution* **70,** 7–17 (2016).
48. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7,** 111–118 (2010).
49. Li, C., Hofreiter, M., Straube, N., Corrigan, S. & Naylor, G. J. Capturing protein-coding genes across highly divergent species. *BioTechniques* **54,** 321–326 (2013).
50. Dimmick, W. W. & Larson, A. A molecular and morphological perspective on the phylogenetic relationships of the otophysan fishes. *Mol. Phylog. Evol.* **6,** 120–133 (1996).
51. Alves-Gomes, J. A. in *Gonorynchiformes and Ostariophysan Relationships* (eds Grande, T., Potayo-Ariza, F. J. & Diogo, R.) (Science Publishers, 2010) 517–565.
52. Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl Acad. Sci. USA* **109,** 13698–13703 (2012).
53. Lavoue, S. *et al.* Molecular systematics of the gonorynchiform fishes (Teleostei) based on whole mitogenome sequences: Implications for higher-level relationships within the Otocephala. *Mol. Phylog. Evol.* **37,** 165–177 (2005).
54. Betancur-R., R. *et al.* The tree of life and a new classificaion of bony fishes. *PLoS Curr. Tree of Life* http://dx.doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288 (2013).
55. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
56. Li, C., Orti, G., Zhang, G. & Lu, G. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol. Biol.* **7,** 44 (2007).

57. Dettai, A. & Lecointre, G. New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii: Teleostei). *Mol. Phylog. Evol.* **48,** 258–269 (2008).
58. Li, C., Riethoven, J. J. & Naylor, G. J. P. EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. *Mol. Ecol. Resour.* **12,** 967–971 (2012).
59. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38,** 7–13 (2010).
60. Sharma, P. P. *et al.* Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol.* **31,** 2963–2984 (2014).
61. Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K. & Nishida, M. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc. Natl Acad. Sci. USA* **112,** 14918–14923 (2015).
62. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
63. R Development Core Team, *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2011); http://www.R-project.org/
64. Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28,** 2685–2686 (2012).
65. Murray, K., Müller, S. & Turlach, B. A. Revisiting fitting monotone polynomials to data. *Comp. Stat.* **28,** 1989–2005 (2013).
66. Maddison, W. & Maddision, D. Mesquite: a modular system for evolutionary analysis, version 2.6. (2009).
67. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).
68. Aberer, A. J., Kobert, K. & Stamatakis, A. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.* **31,** 2553–2556 (2014).
69. Tracer v1.6 (2014); http://beast.bio.ed.ac.uk/Tracer
70. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24,** 774–786 (2008).
71. Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* **58,** 468–477 (2009).
72. Liu, L. & Yu, L. Estimating species trees from unrooted gene trees. *Syst. Biol.* **60,** 661–667 (2011).
73. Shaw, T. I., Ruan, Z., Glenn, T. C. & Liu, L. STRAW: Species TRee Analysis Web server. *Nucleic Acids Res.* **41,** 238–241 (2013).
74. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9,** e1000602 (2011).
75. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent modeL. *Proc. Natl Acad. Sci. USA* **2012** (2012).

Acknowledgements

We dedicate this contribution in honour and memory of our friend and valued colleague Richard Vari whose untimely death has left a huge lacuna in the world of otophysan systematics. We thank D. Maddison, for helping with the MDS analyses in Mesquite, and R. Rivero, for helping with illustrations. We also thank S. Edwards and T. Warnow for providing extensive comments on earlier versions of the paper. J. P. Sullivan kindly provided a photograph for Citharinoidei. This work was supported by National Science Foundation (NSF) grants (DEB-147184, DEB-1541491) to R.B.R., (DEB-1457426 and DEB-1541554) to G.O., (DEB-0315963 and DEB-1023403) to J.W.A., and (DEB-1350474) to L.J.R. This project was also funded by the Opportunity Research Program between George Washington University and the Natural History Museum (Smithsonian) to G.O. and R.V and the Smithsonian Peter Buck fellowship to R.B.R.

Author contributions

D.A., R.B.R., R.V. and G.O. planned the project; R.B.R., K.K and G.O. conducted the pilot experiment; D.A. and R.B.R. carried out the experiments and collected the data; D.A., R.B.R., L.J.R., and G.O. conceived the GGI method; D.A. and R.B.R. analysed data; J.W.A., J.L., M.L.J.S., and M.H.S. collected, identified and curated the fish specimens examined; R.B.R., D.A. and G.O. wrote the paper and all other authors contributed to the writing.

Additional information

Supplementary information is available in the available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.B.R.

How to cite this article: Arcila, D. *et al.* Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat. Ecol. Evol.* **1,** 0020 (2017).

Competing interests

The authors declare no competing financial interests.